# WORK DONE

## 1)Data Ingestion

Deploy

Data Ingestion
Exploratory Data Analysis
Data Cleaning
Feature Importance
Feature Scaling
Model Selection
Model Tuning
Ensembling
Report

### Upload Data

Upload your CSV file

Drag and drop file here
Limit 200MB per file • CSV

Browse files

diabetes.csv 23.3KB

Data preview:

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 2 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 3 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 4 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

# 2. Exploratory Data Analysis

- Data Ingestion
- **Exploratory Data Analysis**
- Data Cleaning
- Feature Importance
- Feature Scaling
- Model Selection
- Model Tuning
- Ensembling
- Report

## Exploratory Data Analysis

☑ Show Descriptive Statistics

### Data Description

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| count | 768 | 768 | 768 | 768 | 768 | 768 | 768 | 768 | 768 |
| mean | 3.8451 | 120.8945 | 69.1055 | 20.5365 | 79.7995 | 31.9926 | 0.4719 | 33.2409 | 0.349 |
| std | 3.3696 | 31.9726 | 19.3558 | 15.9522 | 115.244 | 7.8842 | 0.3313 | 11.7602 | 0.477 |
| min | 0 | 0 | 0 | 0 | 0 | 0 | 0.078 | 21 | 0 |
| 25% | 1 | 99 | 62 | 0 | 0 | 27.3 | 0.2438 | 24 | 0 |
| 50% | 3 | 117 | 72 | 23 | 30.5 | 32 | 0.3725 | 29 | 0 |
| 75% | 6 | 140.25 | 80 | 32 | 127.25 | 36.6 | 0.6263 | 41 | 1 |
| max | 17 | 199 | 122 | 99 | 846 | 67.1 | 2.42 | 81 | 1 |

### Data Types and Null Values

| | Data Type | Null Values | Non-null Count |
|---|---|---|---|
| Pregnancies | int64 | No null values | 768 |
| Glucose | int64 | No null values | 768 |
| BloodPressure | int64 | No null values | 768 |
| SkinThickness | int64 | No null values | 768 |
| Insulin | int64 | No null values | 768 |
| BMI | float64 | No null values | 768 |
| DiabetesPedigre | float64 | No null values | 768 |
| Age | int64 | No null values | 768 |
| Outcome | int64 | No null values | 768 |

# 2. Exploratory Data Analysis

Deploy

- Data Ingestion
- **Exploratory Data Analysis**
- Data Cleaning
- Feature Importance
- Feature Scaling
- Model Selection
- Model Tuning
- Ensembling
- Report

## Highly Correlated Features

| | Feature 1 | Feature 2 | Correlation |
|---|---|---|---|
| 0 | Pregnancies | Age | 0.5443 |
| 1 | Glucose | Outcome | 0.4666 |
| 2 | Insulin | SkinThickness | 0.4368 |
| 3 | BMI | SkinThickness | 0.3926 |
| 4 | Glucose | Insulin | 0.3314 |
| 5 | BMI | Outcome | 0.2927 |
| 6 | BloodPressure | BMI | 0.2818 |
| 7 | Glucose | Age | 0.2635 |
| 8 | BloodPressure | Age | 0.2395 |
| 9 | Age | Outcome | 0.2384 |

## Interpretation of Correlations

**Understanding Correlation:**

Correlation values range from **-1** to **1**:

- **Positive Correlation** (closer to 1): As one feature increases, the other feature tends to increase. Example: Higher study hours leading to better grades.
- **Negative Correlation** (closer to -1): As one feature increases, the other feature tends to decrease. Example: More exercise might result in lower body fat percentage.
- **No Correlation** (closer to 0): Minimal or no linear relationship between features. Example: Shoe size vs. exam scores.

**Importance for Predictive Modeling:**

- Strong correlations (values near ±1) indicate a significant relationship and are often key features for prediction.
- High correlation between independent features can lead to multicollinearity, which may require addressing by removing or combining features to avoid redundancy and overfitting.

# 2. Exploratory Data Analysis

- Data Ingestion
- **Exploratory Data Analysis**
- Data Cleaning
- Feature Importance
- Feature Scaling
- Model Selection
- Model Tuning
- Ensembling
- Report

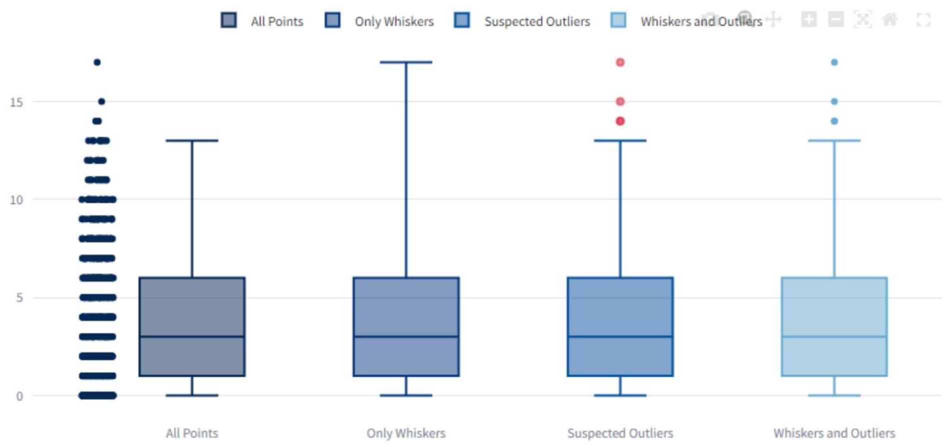
☑ Outlier Investigation

## Single Feature Outliers

Select feature for outlier detection

Pregnancies

### Boxplot for : Pregnancies




### Understanding Box Plots and Outliers

- **Box**: Represents the interquartile range (IQR), which contains the middle 50% of the data.Lower Edge: 1st Quartile (Q1). Upper Edge: 3rd Quartile (Q3). Horizontal Line inside the Box: Median
- **Whiskers**: The lines that extend from the box to the smallest and largest values within 1.5 * IQR.
- **Outliers**: Data points that lie outside the whisker range are considered outliers and are displayed as individual dots.
- **All Points**: Every data point, including outliers.
- **Only Whiskers**: Displays only the key data range within the whiskers, hiding outliers.
- **Suspected Outliers**: Data points that fall outside 1.5 * IQR but aren't extreme enough to be definite outliers.
- **Whiskers and Outliers**: Displays both the whiskers and any definite outliers.

# 3. Data Cleaning

Deploy

- Data Ingestion
- Exploratory Data Analysis
- **Data Cleaning**
- Feature Importance
- Feature Scaling
- Model Selection
- Model Tuning
- Ensembling
- Report

## Data Cleaning

### Outlier Detection Methods

Outlier detection helps identify extreme values in the dataset. Below are two common methods:

1. **Tukey's Method**: Uses the interquartile range (IQR) to detect outliers. It is robust to extreme values and identifies outliers outside the range [Q1 - 1.5*IQR, Q3 + 1.5*IQR].
2. **Z-score Method**: Identifies outliers based on how many standard deviations a data point is from the mean. A common threshold is 3, meaning points more than 3 standard deviations away from the mean are flagged as outliers.

☑ Remove Outliers

Select features to remove outliers from

Pregnancies ×

Select outlier removal method

Tukey's Method

Total number of outliers removed: 4

New dataset has 764 samples and 9 features.

Feature 'Pregnancies': 4 outliers found using Tukey's Method.

☑ Handle Missing Values

No missing values found in the dataset.

☑ Data Type Conversion

Select features for data type conversion

Choose an option

# 4. Feature Engineering

Data Ingestion
Exploratory Data Analysis
Data Cleaning
**Feature Importance**
Feature Scaling
Model Selection
Model Tuning
Ensembling
Report

Deploy

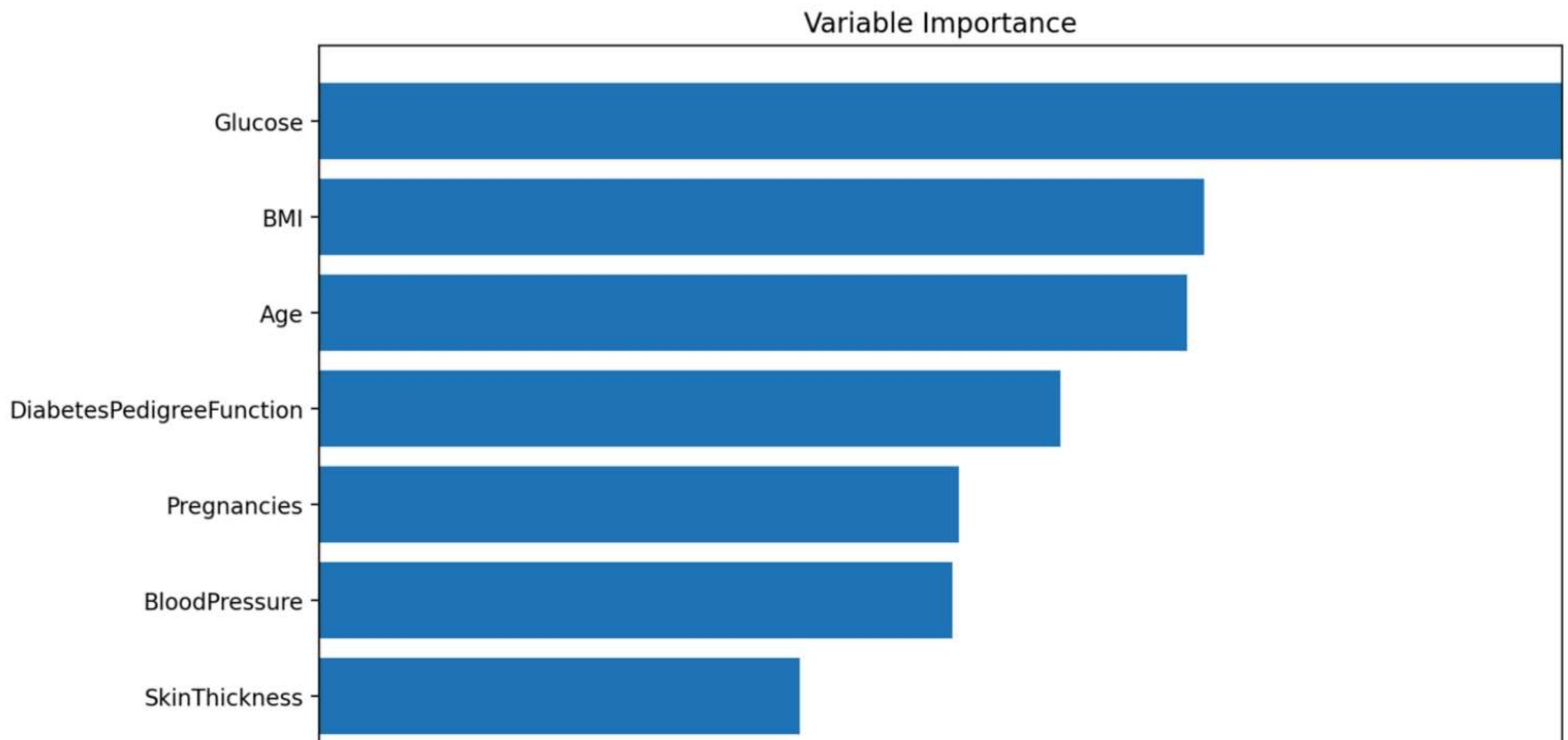
## Feature Engineering

### Select Target Variable

Select the target variable:

Outcome

Independent features: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age



# 5. Feature Scaling

Deploy

- Data Ingestion
- Exploratory Data Analysis
- Data Cleaning
- Feature Importance
- **Feature Scaling**
- Model Selection
- Model Tuning
- Ensembling
- Report

## Feature Scaling / Normalization

Scaling will be applied to the following features: Age, BMI, Glucose

The target variable is: Outcome (which will not be scaled).

Choose a scaling method:

- (●) MinMaxScaler
- ( ) StandardScaler

MinMaxScaler scales the features to a fixed range, usually [0,1].

Scaled Data Preview (including only selected features and target variable):

| | Age | BMI | Glucose | Outcome |
|---|---|---|---|---|
| 0 | 0.4833 | 0.5007 | 0.7437 | 1 |
| 1 | 0.1667 | 0.3964 | 0.4271 | 0 |
| 2 | 0.1833 | 0.3472 | 0.9196 | 1 |
| 3 | 0 | 0.4188 | 0.4472 | 0 |
| 4 | 0.2 | 0.6423 | 0.6884 | 1 |
| 5 | 0.15 | 0.3815 | 0.5829 | 0 |
| 6 | 0.0833 | 0.462 | 0.392 | 1 |
| 7 | 0.1333 | 0.5261 | 0.5779 | 0 |
| 8 | 0.5333 | 0.4545 | 0.9899 | 1 |
| 9 | 0.55 | 0 | 0.6281 | 1 |

Download Scaled Data CSV

# 6. Model Selection

Data Ingestion
Exploratory Data Analysis
Data Cleaning
Feature Importance
Feature Scaling
**Model Selection**
Model Tuning
Ensembling
Report

Deploy

## Model Selection & Baseline Algorithm Evaluation

Model will be trained on features: Age, BMI, Glucose

The target variable is: Outcome

### Binary Classification

The target variable `Outcome` has exactly two unique values, indicating a binary outcome. Therefore, the problem can be modeled as a binary classification task, where the goal is to predict whether an individual falls into one of two categories.

### Suggested Models for Binary Classification

The following models are suggested for binary classification tasks:

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machine
- Decision Tree
- AdaBoost
- Gradient Boosting
- Random Forest
- Extra Trees

### Metric Comparison Across Models

| Model | Accuracy | Precision | Recall | F1-score | ROC AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.8182 | 0.8421 | 0.5926 | 0.6957 | 0.8452 |
| K-Nearest Neighbors | 0.7792 | 0.7273 | 0.5926 | 0.6531 | 0.807 |
| Support Vector Machine | 0.7922 | 0.7895 | 0.5556 | 0.6522 | 0.877 |
| Decision Tree | 0.6753 | 0.5333 | 0.5926 | 0.5614 | 0.6563 |
| AdaBoost | 0.8312 | 0.8889 | 0.5926 | 0.7111 | 0.8893 |
| Gradient Boosting | 0.7922 | 0.7619 | 0.5926 | 0.6667 | 0.8681 |
| Random Forest | 0.7532 | 0.7 | 0.5185 | 0.5957 | 0.8352 |
| Extra Trees | 0.7792 | 0.75 | 0.5556 | 0.6383 | 0.8189 |

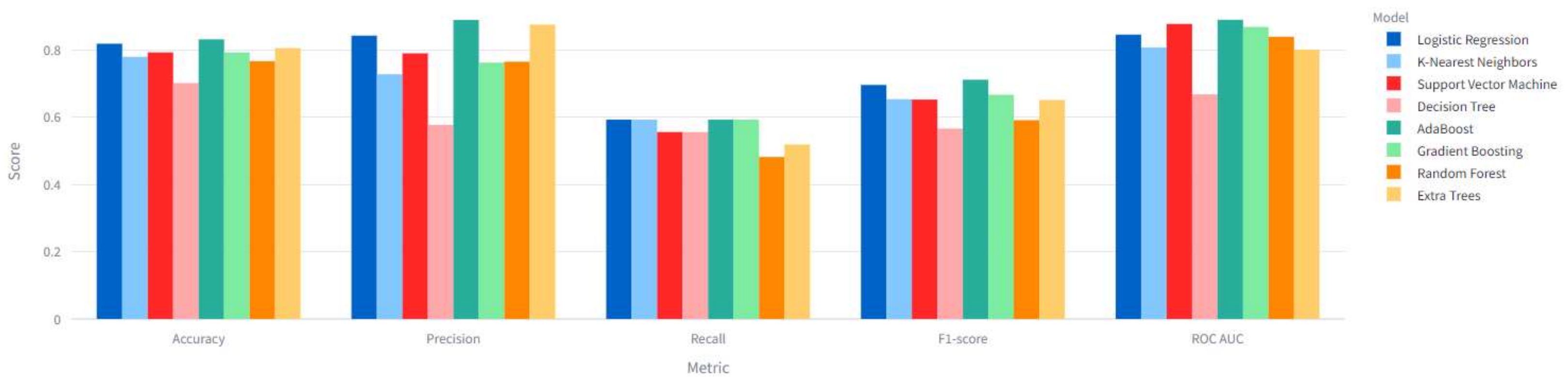
# 6. Model Selection

Deploy

Data Ingestion
Exploratory Data Analysis
Data Cleaning
Feature Importance
Feature Scaling
**Model Selection**
Model Tuning
Ensembling
Report

## Comparison Plot of Model Performance

**Model Performance Comparison**




## Select Models for Ensembling

Choose one or more models for further evaluation:

Logistic Regression × K-Nearest Neigh... × Support Vector ... ×

Selected models saved: Logistic Regression, K-Nearest Neighbors, Support Vector Machine

You have selected the following models for further evaluation:

Logistic Regression, K-Nearest Neighbors, Support Vector Machine

# 7. Model Tuning

Deploy

- Data Ingestion
- Exploratory Data Analysis
- Data Cleaning
- Feature Importance
- Feature Scaling
- Model Selection
- **Model Tuning**
- Ensembling
- Report

## Selected Models for Hyperparameter Tuning

```
[
  0 : "Logistic Regression"
  1 : "K-Nearest Neighbors"
  2 : "Support Vector Machine"
]
```

## Hyperparameter Tuning Methods

- [x] Auto Tuning Hyperparameters
- [ ] Manual Tuning Hyperparameters

## Auto-Tuning Hyperparameters

Tuning: Logistic Regression

Tuning: K-Nearest Neighbors

Tuning: Support Vector Machine

## Best Hyperparameters for Each Model:

Logistic Regression: {'penalty': 'l2', 'C': 0.1}

K-Nearest Neighbors: {'n_neighbors': 7, 'weights': 'uniform'}

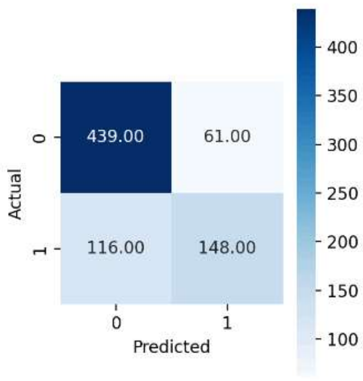
Support Vector Machine: {'C': 10, 'kernel': 'poly'}

Deploy

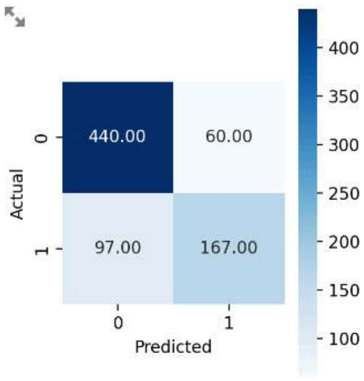
- Data Ingestion
- Exploratory Data Analysis
- Data Cleaning
- Feature Importance
- Feature Scaling
- Model Selection
- **Model Tuning**
- Ensembling
- Report

# Evaluation Metrics for Tuned Models:

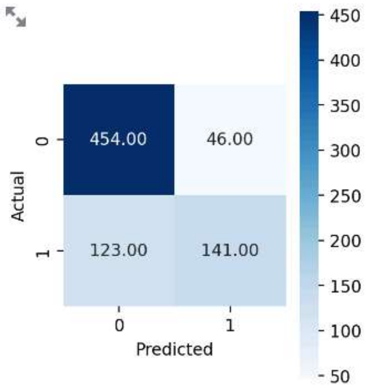
## Logistic Regression Confusion Matrix




## K-Nearest Neighbors Confusion Matrix




## Support Vector Machine Confusion Matrix




# Classification Reports:

## Classification Report for Logistic Regression:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.7910 | 0.8780 | 0.8322 | 500.0000 |
| 1 | 0.7081 | 0.5606 | 0.6258 | 264.0000 |
| accuracy | 0.7683 | 0.7683 | 0.7683 | 0.7683 |
| macro avg | 0.7496 | 0.7193 | 0.7290 | 764.0000 |
| weighted avg | 0.7624 | 0.7683 | 0.7609 | 764.0000 |

# 8.Ensembling

Deploy

- Data Ingestion
- Exploratory Data Analysis
- Data Cleaning
- Feature Importance
- Feature Scaling
- Model Selection
- Model Tuning
- **Ensembling**
- Report

## Super Learner Training

☑ Select All Models

Selected models:

[
0 : "Logistic Regression"
1 : "K-Nearest Neighbors"
2 : "Support Vector Machine"
]

Train Super Learner

Super Learner trained successfully!

Super Learner Accuracy: 0.8438

Super Learner Precision: 0.8571

Super Learner Recall: 0.6000

Super Learner F1-score: 0.7059

Super Learner ROC AUC: 0.7773